题目自动生成的技术演进与质量控制*

韩雨婷 1,2,3　王文轩 4　刘红云 5,6　游晓锋 7

([1] 北京语言大学心理与认知科学学院) ([2] 北京语言大学生命健康研究院)
([3] 语言认知科学教育部重点实验室，北京 100083)
([4] 香港科技大学计算机科学与工程系，香港 999077)
([5] 应用实验心理北京市重点实验室) ([6] 北京师范大学心理学部，北京 100875)
([7] 南昌师范学院数学与信息科学学院，南昌 360111)

摘 要　题目自动生成（Automatic Item Generation，AIG）技术通过自动化生成测验题目，旨在解决心理与教育测验中题目开发成本高、效率低、维护困难和安全风险等问题。但在提高效率的同时，如何保障题目质量仍是关键挑战。为此，梳理了 AIG 的理论基础，分析了从基于规则到数据驱动的技术演进历程，系统考察了不同类型测验中的应用实践，探讨了质量控制的多层次保障机制。提出了认知理论与深度学习融合、知识图谱与检索增强生成技术、提示工程优化、多模态技术融合和多层次质量评估等改进路径，以推动 AIG 从单一工具向成熟的智能测验系统转变，全面提升自动生成题目的质量与可靠性。

关键词　题目自动生成；质量控制；心理测量学；大语言模型

1 引言

心理与教育测验作为评估个体心理特征、能力水平和行为表现的重要工具，在基础研究、临床诊断、教育教学、人才选拔等领域发挥着关键作用（Hambleton, 2004）。随着心理测验在研究和实践中的广泛应用，对测验题目的质量和数量需求急剧增长。特别是计算机化适应性测验（Computerized Adaptive Testing, CAT）的推广，要求建立包含大量等值题目的题库以保证测验的信度和效度（Embretson & Yang, 2007）。然而，传统的人工编制题目方式存在显著局限。首先，命题专家需要投入大量时间和精力进行题目编写、评审和修订（Kurdi et al., 2020）。其次，题目资源有限，在高利害测验中容易出现题目被记忆和传播的问题，严重影响测验的安全性（Bejar et al., 2003）。此外，人工编制难免受到命题者主观经验和认知偏差的影响，难以保证题目质量的一致性（王蕾等, 2023）。因此，如何高效地生成质量高、数量充足的测验题目，已成为心理与教育测量领域的重要挑战。

题目自动生成（Automatic Item Generation, AIG）是指计算机根据特定要求，基于认知理论和心理测量模型，自动生成符合指定参数的测验题目（李中权和张厚粲, 2008；Embretson & Yang,

---

* 通信作者:刘红云, E-mail: hyliu@bnu.edu.cn

2007）。作为认知心理学、心理测量学和人工智能技术的交叉产物，AIG 具有以下显著特征：能够生成指定难度水平和具备合适心理测量学特性的题目；可实现无需试测直接使用；具备题目层面的结构效度（Embretson, 2002）。更重要的是，AIG 系统能够快速批量生成大量题目，即使是半自动生成系统，其题目生成效率也会显著提升（Mitkov et al., 2006）。

AIG 技术的发展经历了多个重要阶段（Haladyna, 2013）。侧面理论（Facet Theory）旨在通过映射句子（包含固定和可变部分）来操作性定义测验内容并进行统计验证，其中映射句子由学科专家创建，而干扰项则通过算法自动生成（Guttman, 1953; 1959）。基于文本的自动命题方法（Prose-based AIG）尝试从优质散文中提取关键句子并转换为问题形式（Bormuth，1970）。此外，Hively（1974）提出了基于固定句法结构的题目生成技术，进一步推动了该领域的发展。20 世纪 60 至 80 年代，概念学习成为 K-12 教育和培训的主要目标，研究者将重点转向概念测量。1998 年，教育考试服务中心（ETS）举办的题目自动生成研讨会及其促成的《Item Generation for Test Development》一书的出版，标志着该领域进入理论成熟阶段。

在 AIG 技术实现层面，早期主要形成了两种方法路径（Embretson & Yang, 2007；Bejar et al., 2003）：一是基于强理论的认知设计系统法（Cognitive Design System），通过分析问题解决过程中隐含的心理学原理，精细控制测验模型的难度；二是基于弱理论的题目模型法（Item Modeling Approach），以一组具有内容和难度代表性的校准题目为基础进行生成。这两种方法为后续研究奠定了重要基础。随着机器学习，特别是大语言模型（Large Language Models，LLMs）的快速发展，AIG 技术获得了新的突破。基于深度学习的方法不仅显著提高了生成效率，还在一定程度上克服了传统方法中模板化、缺乏创新性等局限（Götz et al., 2023）。近年来，以 ChatGPT 为代表的人工智能生成内容（Artificial Intelligence Generated Content， AIGC）技术的突破性进展，为 AIG 的发展开辟了新的技术路径，推动 AIG 向更智能、更灵活的方向迈进。同时，研究者开始探索将技术增强型试题（Technology-Enhanced Items）与自动生成相结合，以实现更复杂的能力测评（王蕾等, 2023），预示着 AIG 技术正朝着智能化和个性化的方向发展。

虽然 LLMs 等人工智能技术为题目生成提供了创新思路，但在专业知识表达的准确性、构念效度的保障以及测量质量的稳定性等方面仍面临重要挑战。这些挑战不仅涉及技术层面的突破，更需要深入理解测验编制的理论基础和质量要求。为此，本研究聚焦于 AIG 领域的一个关键科学问题：如何在提高题目生成效率的同时，确保自动生成题目的科学性与测量质量。本研究将系统梳理 AIG 技术的理论体系、技术演进路径、应用实践和质量控制机制，揭示影响自动生成题目质量的关键因素。在此基础上，还将深入探讨当前 AIG 面临的挑战和未来发展方向。研究成果将为提升心理与教育测验 AIG 的科学性与实用性提供理论指导、技术方案及新思路，推动心理与教育测验向更高效、更精准的现代化方向发展。

2 题目自动生成的理论基础

作为多学科交叉的产物，AIG 主要涉及认知心理学、心理测量学、语言学和计算机科学。这些学科的理论相互支撑，共同构建了 AIG 的理论体系。本节将从这四个维度系统阐述其理论框架及其关联，为提升自动生成题目的质量提供理论指导。

2.1 认知心理学理论基础

认知心理学为 AIG 提供了理解解题过程、认知机制和难度控制的理论基础。

认知负荷理论（Cognitive Load Theory，Sweller，1988）强调了个体在信息加工过程中认知资源的有限性。该理论指出，学习者的工作记忆容量有限，认知负荷的大小直接影响学习效果。在 AIG 中，认知负荷理论为控制题目难度提供了理论依据。通过调整题目的信息量、复杂度和干扰项，AIG 可以控制学习者在解题时的认知负荷。例如，在设计阅读理解题目时，控制文本的句法复杂度、生词密度等，可以调整题目的难度水平（Embretson & Yang, 2007）。

信息加工理论（Information Processing Theory，Atkinson & Shiffrin, 1968）将人类认知过程视为信息输入、编码、存储和提取的过程。该理论为理解学习者在解题时的信息处理方式提供了框架。在 AIG 中，信息加工理论指导我们分析题目所需的认知过程，包括感知、注意、记忆和思维等环节。例如，在数学题目生成中，需要考虑解题者对数学概念的理解、问题情境的认知，以及解决问题所需的推理过程。

Embretson（1998）提出的认知加工模型（Cognitive Processing Models）将认知心理学与心理测量学相结合，用于分析题目解决过程中涉及的认知成分。该模型强调通过识别影响题目难度的基本成分（radicals）和随机成分（incidentals），可以有效地控制题目难度。基本成分是对题目难度有显著影响的特征，如工作记忆需求、认知操作复杂度等；随机成分则是不影响题目难度的表层特征，如题目中的情境描述等。在 AIG 中，可以根据认知加工模型精细地设计和生成难度可控的题目。

2.2 心理测量学理论基础

心理测量学为 AIG 提供了效度验证、题目与测验参数评估的理论基础。

构念效度（Construct Validity）是评估测验是否有效测量目标心理特质的关键概念。Embretson（1983）将构念效度分为构念表征和规则广度两个维度。构念表征关注鉴别任务表现潜在的认知成分。在 AIG 中，需要通过构念表征建立认知加工模型，指导题目的自动生成。例如，在矩阵推理测验中，通过认知模型分析确定规则数量和复杂度等基本成分如何影响题目难度；在人格测验中，则需要明确题目内容如何准确反映特定人格维度的理论内涵。规则广度指

测验分数与其他测量之间的相关网络，体现在与其他特质测验、效标测量等的关系强度、频率和模式上。例如，自动生成的题目应保持与原有测验相似的测量特性，其分数与其他特质测验、效标测验等的关系模式应保持稳定。这要求 AIG 系统不仅要能将构念表征的理论要求转化为题目生成规则，还要确保生成的题目能够保持原有测验的规则广度特征。

项目反应理论（Item Response Theory，IRT）通过建立考生能力水平、题目特征参数（如难度、区分度）与作答反应之间的数学模型，为题目参数的估计和预测提供了工具（Embretson & Reise, 2000）。在 AIG 中，IRT 可用于评估自动生成题目的心理测量学特性，确保其具有良好的信度和效度。

线性逻辑模型（Linear Logistic Test Model，LLTM，Fischer, 1973）是 IRT 的扩展，允许将题目难度分解为若干可解释的认知操作成分的线性组合。LLTM 使得基于认知特征预测题目难度成为可能。例如，在数学应用题中，LLTM 可以分析编码难度、解题步骤数量等具体认知成分对题目难度的影响（Embretson & Daniel, 2008）。

2.3 语言学理论基础

语言学理论为题目自动生成提供了表达和理解层面的基础支持。虽然这些理论最初主要应用于语言测试领域，但其核心观点对所有类型的题目生成都具有重要的指导意义。

Bachman（1991）提出的交际语言能力模型深化了对语言能力构念的认识。其中“真实性”这一核心概念涵盖了情景和交际两个维度。Bachman 和 Palmer （1996）将真实性定义为测试任务特征与实际任务特征的一致性程度，这一定义不仅对语言测试有指导意义，也为其他学科领域题目的情境设计和质量评价提供了重要参考。例如，在生成数学应用题时，应选择贴近学生日常生活的情境，而不是使用过于抽象或脱离实际的问题描述。

Larsen-Freeman（1997）将动态系统理论引入应用语言学，认为语言是一个复杂的动态自适应系统，不仅其中语音、词汇、语义等子系统相互依赖、密切关联，还会随社会发展不断演变。这对 AIG 的启示在于：生成题目时需要统筹考虑语言的多个组成部分,确保在词汇选择、语法结构和表达方式等方面的协调统一；同时还要反映语言使用的动态特征，注意词汇语义系统的实时更新，以保持测验内容的时代适切性。

图式理论（Rumelhart, 1980； Anderson & Pearson, 1984）、理解层次模型（Van Dijk & Kintsch, 1983）和系统功能语言学理论（Halliday & Matthiessen, 2013）为题目难度控制提供了重要的理论基础。图式理论指出读者对文本的理解建立在已有知识结构（图式）之上。这些图式包括内容知识、文本组织方式的认识以及语言知识。当文本内容与读者已有图式越接近，理解就越容易。理解层次模型则揭示了文本理解包含表层结构（词句字面意思）、文本基础（命题间关系）和情境模型（与已有知识整合）等多个层次，不同层次需要不同的认知加工深度。系统功能语

言学理论强调语言具有概念、人际和语篇三大功能，这些功能都嵌入在特定的社会语境中。基于这些理论，题目难度可通过多个维度调控：调整主题与已有知识的关联度（如低年级用校园生活，高年级用太空探索），设置不同层次的理解要求，以及根据语言发展水平调整表达方式（如低年级用具体指导语，高年级用抽象主题），从而实现对认知和语言发展水平的科学测量。

2.4 计算机科学理论基础

计算机科学，特别是自然语言处理和深度学习，为 AIG 的实现提供了强有力的技术支持。

自然语言处理（Natural Language Processing，NLP）旨在使计算机理解和生成人类语言（Jurafsky & Martin, 2022）。在 AIG 中，NLP 技术用于自动生成题目文本、分析语法结构、识别语义关系等。例如，通过词性标注和句法分析可以识别文本的关键成分，进而生成填空题；通过语义理解和语义相似度计算可以生成合理的干扰项。

深度学习（Deep Learning）的发展，特别是神经网络模型的引入，显著提升了 NLP 的性能。序列到序列（Sequence-to-Sequence）模型和注意力机制（Attention Mechanism）进一步提升了文本生成的连贯性和流畅度。特别是 Transformer 架构（Vaswani et al., 2017）通过自注意力机制有效解决了长文本处理中的长距离依赖问题，为题目生成提供了更好的语义理解和表达能力。预训练语言模型（Pre-trained Language Models），如 BERT（Devlin et al., 2019）和 T5（Raffel et al., 2020），以及发展至今的大语言模型如 GPT 系列（Brown et al., 2020），通过在大规模文本数据上预训练，学习了丰富的语言表示和知识，获得了强大的语言理解和生成能力。这些模型为 AIG 的发展提供了新的可能性，研究者们正在探索将其应用于题目生成、答案评估等任务。例如，美国医学考试委员会利用 GPT-2 模型成功生成了医疗认证题目（von Davier, 2019）。

预训练语言模型通过有监督微调（Supervised Fine-tuning）和提示工程（Prompt Engineering）可以更好地适应特定任务需求。有监督微调通过在特定任务数据集上进行训练，使模型学习任务相关的知识和模式。但这一过程会改变模型的参数，需要较大的计算资源和专业技术支持。相比之下，提示工程（Prompt Engineering）是一种更轻量级的方法，它不需要改变模型参数，而是通过设计合适的提示（Prompt）来引导已有模型生成目标内容（Liu et al., 2021）。在 AIG 中，可以通过提供示例题目、明确任务要求等方式，来提升生成题目的相关性和质量（Götz et al., 2023）。这种方法实施成本较低，且具有较好的灵活性。结合这两种方法的特点，研究者可以根据实际需求和资源条件，选择合适的方式来优化 LLMs 在 AIG 领域的应用。

2.5 理论基础小结

综上，AIG 的理论基础涵盖了认知心理学、心理测量学、语言学和计算机科学等多个领域。认知心理学通过任务分解和认知成分分析，为理解解题过程、控制题目难度提供了基础；心理

测量学为题目的质量评估和测量效度提供了框架；语言学理论为题目的语义表达和语境构建提供了规范；计算机科学的 NLP 和深度学习理论则为 AIG 的实现提供了技术支持。在具体应用中，AIG 仍面临一些挑战。例如，如何准确建模题目难度与认知成分的关系，如何确保生成内容的专业性和准确性，以及如何避免 LLMs 生成的不可靠内容（幻觉现象）等。未来的研究方向包括深化认知模型的构建，完善题目参数预测模型，优化提示工程方法，以及加强多学科的融合与协作。

3 题目自动生成的技术演进

在追求生成题目的科学性与测量质量的过程中，技术方法选择起着关键作用。从早期的规则驱动方法，到语料库方法，再到当前融合深度学习与 LLMs 的智能生成方法，每次技术演进都在效率与质量间寻求平衡。这些方法虽然都致力于解决质量控制问题，但在实现路径和效果上差异显著。本节将系统分析这些方法的理论原理、实现机制及其在保障测量质量方面的优劣，以探索技术创新促进题目生成科学性的路径。

3.1 基于规则的方法

基于规则的方法是 AIG 研究中最早应用的技术，主要依赖于专家知识和认知理论，通过预先定义的规则或模板来控制题目的生成过程。根据设计策略的不同，主要包括认知设计系统法、题目建模法和基于本体的方法。这三种方法的具体流程如图 1 所示。

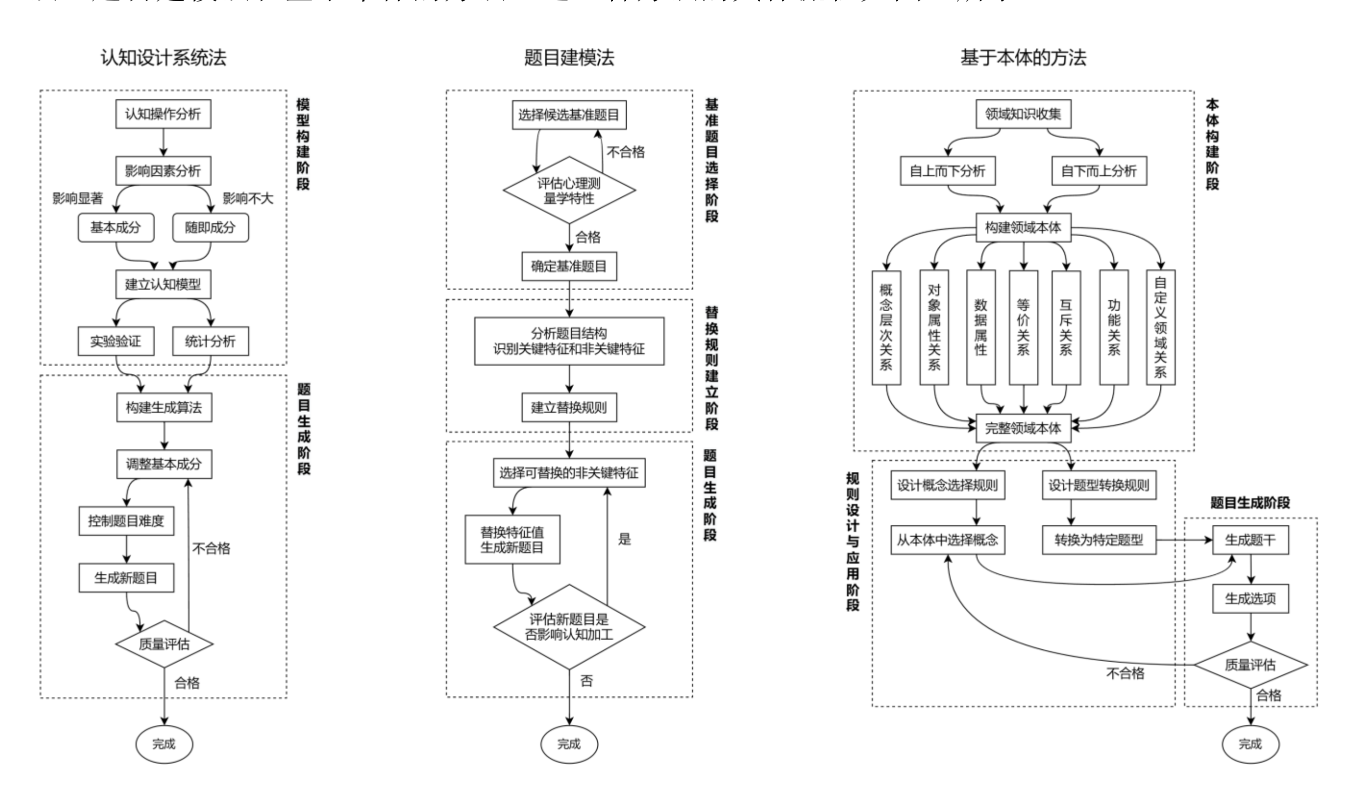


图 1 基于规则的题目自动生成流程图

3.1.1 认知设计系统法

认知设计系统法（Cognitive Design System Approach）的基本思想是通过对题目解决过程中的认知操作进行分析，识别影响题目难度的关键因素，即基本成分和随机成分（Embretson, 1998）。基本成分是对项目心理测量学特性（如难度、区分度）有显著影响的刺激特征，随机成分则对项目特性影响不大。

在具体实施中，研究者首先建立现有题目的认知模型，采用实验和统计方法验证各刺激特征对题目难度的影响（Embretson & Yang, 2007）。然后，根据认知模型构建题目生成算法，通过调整基本成分来控制题目难度。例如，在矩阵推理测验中，通过改变规则的数量和复杂度来生成不同难度水平的题目（Embretson & Daniel, 2008）。

认知设计系统法的优势在于能够根据认知理论精确控制题目难度，生成具有良好心理测量学特性的项目。但其局限性在于对认知模型的依赖较高，需要深入的理论研究和实验验证，适用于已有坚实认知理论基础的测验类型。

3.1.2 题目建模法

题目建模法（Item Modeling Approach）也称为模板法，以具有良好心理测量学指标的题目为原型，通过替换非关键特征（如人物、情境、数值等）生成新题目（Bejar et al., 2003）。

实施过程包括：首先，选择和评估基准题目，确保其具有良好的心理测量学特性（Gierl et al., 2016）。其次，识别可替换的非关键特征，建立替换规则，注意避免影响题目的认知加工过程。最后，通过替换生成新题目。例如，在数学应用题中，可以通过改变题目中的人物名称、具体数值等，生成等价的题目。

题目建模法的优势在于实施简便，成本较低，适用于需要大量平行测试题目的情况。但其生成的题目变化较小，形式固定，难以产生创新性强的题目。

3.1.3 基于本体的方法

基于本体的方法利用本体论（Ontology）对领域知识进行形式化表示，通过规则生成题目（Papasalouros et al., 2008）。该方法包含两个关键步骤：构建领域本体和设计生成规则。

领域本体通过概念层次、属性和关系等方式表示领域知识（丁向民, 2008），如在构建航空领域的“飞机”本体时，包含概念层次（“飞机”是交通工具的一种）、属性（飞机的“速度”、“航程”等特征）以及关系（“发动机是飞机的组成部分”）等方面。研究者通过自上而下和自下而上相结合的方式来构建这些知识表示。在规则设计方面，包括概念选择（从本体中选择合适的概念作为题目内容）和题型转换（将选择的概念转换为特定类型的题目）两类规则。例如，基于上述本体关系，可以自动生成如“下列哪项不是飞机的基本组成部分？”这样

的选择题，并根据本体中的关系自动生成包含“发动机”在内的正确选项和干扰项。

这种方法能够生成专业性强、逻辑严谨的题目，特别适用于专业学科测验（Alsubait et al., 2014），但本体构建和规则设计需要大量专业知识投入，维护成本较高。

3.2 基于语料库的方法

基于语料库的方法通过分析大规模语料库来生成题目。研究者可以利用现有的语言数据联盟语料库（Linguistic Data Consortium，LDC）、牛津文本档案库（Oxford Text Archive，OTA）等资源，或通过 Web 爬取构建语料库。这种方法的核心是对文本进行挖掘和分析，提取题目生成所需的语义和结构特征。基于语料库的题目生成主要包括三个步骤：语料选择与预处理、题干生成和干扰项生成（Smith et al., 2010）。图 2 展示了基于语料库的 AIG 方法实现过程。

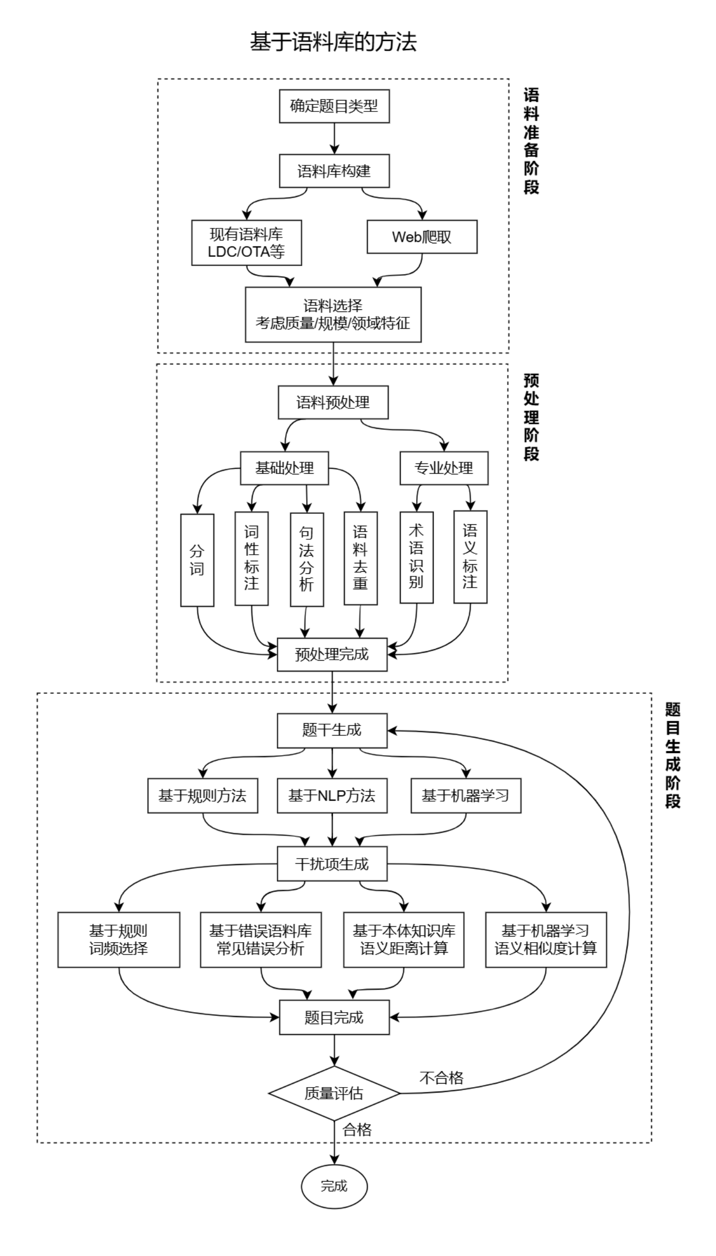


图 2 基于语料库的题目自动生成流程图

3.2.1 语料选择与预处理

语料库的选择主要考虑规模和应用目标。Smith 等人（2010）使用 15 亿词的 UKWaC 语料库证实了大规模语料对提升题目多样性的重要性。Biber 等人（1998, 2002, 2004）通过分析大学课本发现，语料专业性与测试目标的匹配度影响测试效度，这在 Barker（2006）的 ESOL 考试研究中得到验证。此外，Ball（2002）和 Barker（2010）则证实了母语和学习者语料库的互补使用能有效提高测试的科学性。在预处理方面，Mitkov 和 Ha（2003）建立了句法分析、词性标注和术语抽取的技术框架，Karamanis 等人（2006）通过术语识别和过滤技术提升了处理精度。

3.2.2 题干生成

题干生成技术经历了从规则驱动到机器学习的演变。早期 Sumita 等人（2005）通过词形分析和考点匹配确定空白项位置。近期研究转向机器学习方法，如 Mitkov 和 Ha（2003）引入浅层句法分析和术语抽取等 NLP 技术提升了题干结构合理性。Goto 等人（2010）将偏好学习用于句子特征提取，采用条件随机场估计空白项。Smith 等人（2010）则通过融合语料库和规则约束，建立了包含句长、标点等的多维度质量控制体系。

3.2.3 干扰项生成

一些研究者利用语料库数据选择干扰项。如 Sakaguchi 等（2013）从 Lang-8 学习者语料库中抽取错误类型，使用支持向量机训练分类器生成具有混淆性的干扰项。除语料库方法外,研究者还探索了其他技术路径（肖文艳, 2019）：如 Coniam（1997）基于词频规则选择干扰项；Agarwal 和 Mannem（2011）进一步考虑句法和语义特征，从给定文本中选择具有相似语境的词作为干扰项；Mitkov 等人（2006）则利用 WordNet 的语义网络计算词义距离来确定干扰项。这些方法的共同目标是生成与正确答案在形式或语义上相近的干扰项。

3.3 基于深度学习的方法

基于深度学习的 AIG 通过神经网络学习语言特征，能生成更自然流畅、创新性更强的题目。这类方法经历了从词嵌入到大语言模型的发展历程。

3.3.1 基于经典的深度学习方法

早期方法以词嵌入模型（如 Word2Vec）为主，通过计算词向量相似度生成干扰项。研究者随后探索循环神经网络（RNN）和长短期记忆网络（LSTM）的应用，如 von Davier（2018）将 LSTM 用于人格测验题目生成，但这些方法在处理长距离依赖和复杂结构时仍有局限。

3.3.2 基于大语言模型的方法

基于 Transformer 架构的大语言模型通过海量文本预训练获得了强大的语言理解和生成能力。在题目生成中主要通过两种方式应用：一是领域微调，如 Hommel 等人（2022）使用 GPT-2 生成人格测验题目，约 2/3 具备良好测量特性；二是提示工程，通过设计包含题型说明、示例、难度要求等的模板引导生成，如 Attali 等人（2022）利用 GPT-3 生成阅读理解题目及解析。

3.3.3 最新进展

在提示工程方面，研究者探索了多种提升生成质量的策略。Brown 等人（2020）提出的少样本学习（Few-shot Learning）通过提供高质量示例指导模型理解题目结构和难度要求。思维链提示（Chain-of-Thought Prompting）则通过引导模型展示推理过程，使其逐步思考知识点、设计题干和选项（Wei 等, 2022）。Kojima 等人（2022）进一步提出零样本思维链（Zero-shot-CoT）方法，即使无示例也能通过步骤化提示激发模型推理。此外，通过让模型进行角色扮演（Shanahan et al., 2023），如扮演专业教师，可以帮助生成更规范的题目。

研究者还探索了多种混合方法以提升生成质量。高凯（2024）提出结合知识图谱的方法，通过结构化知识表示提供准确的语义和逻辑信息。此外，结合传统 NLP 技术（如关键词提取、句法分析）和规则系统进行后处理优化，也能提高题目规范性。特别值得关注的是检索增强生成（Retrieval-Augmented Generation，RAG）技术的应用。RAG 通过实时检索外部知识库为模型提供专业背景知识，有效解决了传统生成模型在复杂任务中的局限，其基本原理如图 3 所示。在实践中，王鹏等（2024）利用 RAG 技术实现了心理危机评估量表的自动生成，陈欣等（2024）将学科知识导入向量数据库实现知识增强，结合教师细分的课程内容知识点纳入提示模板，调用大语言模型生成题目。这种方法通过知识库构建、检索方案设计和提示模板配置，实现了题目的高效自动生成，同时保证了生成内容的专业性。图 4 展示了基于大语言模型的 AIG 方法框架。

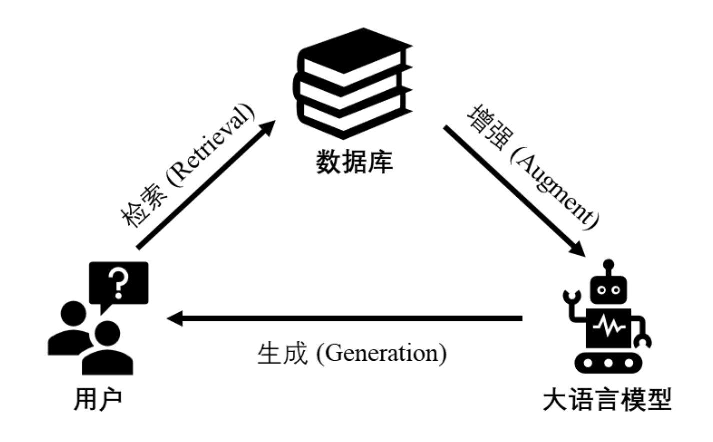


图 3 检索增强生成（RAG）技术的基本原理

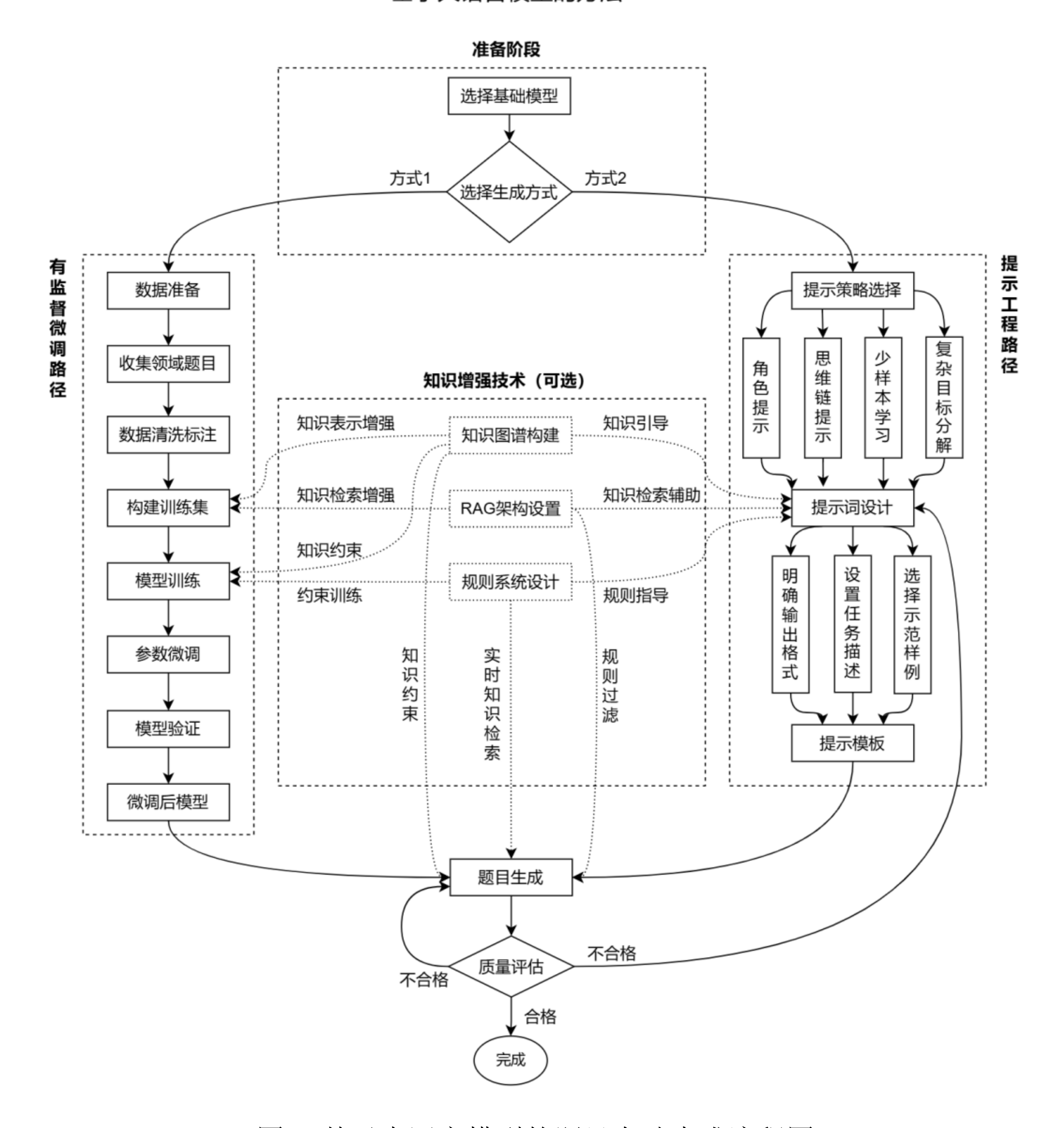


图 4 基于大语言模型的题目自动生成流程图

3.4 技术演进小结

AIG 技术发展历程显示，各技术范式在解决测量质量问题时各有优劣：基于规则的方法控制精确但维护成本高，难以处理复杂语言现象且缺乏多样性；基于语料库的方法能从真实语料中学习语言模式且支持多维度难度匹配，但受限于语料质量，在专业领域面临知识更新不及时问题；基于深度学习的方法效率高、创新性强但质量不稳定，对测验的科学性和公平性构成挑战。为突破这些技术瓶颈，未来研究应发展融合规则约束与深度学习的混合架构，将专业知识图谱、RAG 与 LLMs 结合以确保专业性与准确性，同时通过优化提示工程（如示例学习、思维链提示等）的任务描述和示例设计，建立智能化质量评估机制，实现对生成题目测量特性的快速预测和筛选，从而解决效率与质量平衡这一核心问题

4 题目自动生成的应用实践

由于不同类型的测验具有独特的构念特征和应用要求，其质量控制难点也各不相同。本章从教育、认知、人格和心理健康几个领域出发，系统分析各类测验运用 AIG 技术时的具体质量控制问题及其解决方案，为后续研究指明方向。

4.1 AIG 在教育测验中的应用

4.1.1 数学测验题目自动生成

数学测验题目主要分为应用题和非应用题两类（申弋斌, 2022），分别测评学生的数学建模与实践应用能力，以及函数、几何、代数等学科知识（高凯, 2024）。AIG 技术在数学测验中的应用经历了从规则驱动到智能生成的演进：早期主要采用基于模板的方法，通过填充数值和情境内容生成同构题目（Singley & Bennett, 2002）；随后发展出融合语义理解和知识表示的神经网络模型，如 MAGNET 模型通过语义抽取和实体约束实现了数学应用题的自动生成（Zhou & Huang, 2019）；近期则出现了结合 LLMs 和知识图谱的混合方法，显著提升了生成题目的质量与多样性（高凯, 2024）。然而，数学题目自动生成仍面临三个关键挑战：一是复杂语义关系和隐含条件的准确建模，二是非应用题（如函数、几何）的生成方法有待突破，三是与教学知识体系的深度融合。未来研究应着重完善语义理解模型、扩展题型覆盖范围，并加强与教学目标的关联性，实现更精准的能力测评。

4.1.2 语言能力测验题目自动生成

语言能力测验的自动生成研究涵盖英语、汉语、西班牙语等多种语言，涉及词汇、语法、阅读理解等方面（LaFlair et al., 2023），包含多项选择、词汇判断、完形填空等多种题型（Attali et al., 2022; Sayin et al., 2024）。从技术发展脉络看，语言题目生成经历了三个阶段：早期基于规则和模板的专家系统（Gierl & Haladyna, 2013）；中期利用语料库和统计方法提升真实性，如 Shei（2001）基于词频和上下文分析选取填空位置，Smith 等人（2010）的 TEDDCLOG 完形填空自动生成系统（Smith et al., 2010），以及基于词汇语义网络（Liu et al., 2005）和同义词典（Sumita et al., 2005）的干扰项生成；近期则转向深度学习方法，从 Jiang 和 Lee（2017）的词嵌入模型生成干扰项，到 Du 等人（2017）的端到端问题生成，再到徐坚（2024）的图注意力网络模型，生成质量不断提升。特别是 LLMs 的应用，在题目生成和难度预测方面都展现出显著优势（Sayin et al., 2024）。如 Duolingo 应用 GPT-3 模型实现了交互式阅读理解题目的生成与自动评分（Attali et al., 2022）

语言测验 AIG 仍面临多重挑战：首要问题是高利害测验场景下的质量把控，即使采用先进语言模型，生成结果仍需人工审核（Attali et al., 2022）；其次是干扰项设计的科学性，需要

在语义合理性和迷惑性间寻找平衡（Gates et al., 2011）；此外，生成题目与教学大纲的匹配度（Liu et al., 2005）和难度预测的准确性（肖文艳，2019）都有待提高。特别的，阅读理解测验对真实语料的需求，如何建立兼顾内容多样性、语言真实性和难度分级的自动化筛选机制亟待解决。

4.1.3 专业学科测验题目自动生成

专业学科测验 AIG 的关键挑战在于确保生成内容的专业性和准确性，研究主要集中在医学、法律和生物医学等领域。在医学领域，技术路线经历了显著演进：从 Mitkov 等人（2006）基于浅层句法分析和 UMLS 医学词库的快速生成系统，到 Karamanis 等人（2006）结合 Charniak 解析器、UMLS 词库和分布相似度计算的快速题目生成（RIG）系统，再到 von Davier（2022）利用 GPT-2 模型在 80 万篇 PubMed 文献上微调实现临床病例及题目生成。法律领域则聚焦于专业知识的提取与转化，Zheng 等人（2021）构建了 CaseHOLD 数据集用于案例问答系统训练，Mitkov 等人（2023）通过 doc2vec 和 SBERT 模型实现了法律文本的题目生成。此外，在航空工程和生物医学领域，研究者多采用基于本体的方法，如丁向民（2008）和 Papasalouros 等人（2008）分别构建了航空和生物医学知识本体，通过概念层次和属性关系生成题目，并通过使用同一父类下的其他子类、具有相似属性的其他概念以及利用本体中定义的不相容关系来生成干扰项。

专业测验 AIG 虽已实现从规则驱动向深度学习的转型，但在专业术语准确性、场景真实性和知识关系表达等方面仍有重要挑战。未来研究需要探索将专业领域知识库与深度学习模型深度融合的技术方案，以保障生成题目的专业质量。

4.2 AIG 在基础认知能力测验中的应用

基础认知能力测验 AIG 主要采用两种技术路线：基于模板的方法和基于 LLMs 的方法。Fu 等人（2022）开发了数量和逻辑推理题目模板，通过控制干扰项和图形元素变换（如文氏图的旋转和镜像）生成同构性题目，发现数值范围的变化会显著影响题目难度。Laverghetta 等人（2023）利用 GPT-3 生成自然语言推理测验题目，通过最佳和最差样例引导生成过程，在命题结构（逻辑关系）和量词（数量关系）等特定构念测试中取得了良好的效度。

认知能力测验 AIG 面临构念测量的稳定性与多样性平衡问题，同时在图形、空间等多模态内容的自动生成方面仍有重要挑战。这需要在认知理论指导下，融合模板法、LLMs 和多模态生成技术，构建更全面的认知测验生成框架。

4.3 AIG 在人格测验中的应用

人格测验 AIG 由于构念和语言表达的复杂性起步较晚（von Davier, 2018）。von Davier

（2018）首次使用 LSTM 生成语法和形式正确的题目，但无法针对特定人格特质；Hommel 等人（2022）通过微调 GPT-2 模型并输入构念标签实现了针对性生成（如“Pessimism”生成“I am not likely to succeed in my goals”），约 2/3 的题目具有良好的心理测量特性。近期，Götz 等人（2023）基于 GPT-2 开发的心理测量题目生成器（PIG）通过漫游症题目生成和 Big Five 量表简版（N-BFI-20）的开发验证了其实用性，生成题目展现出与人工编制相当的信效度水平。

人格测验 AIG 仍面临题目质量不稳定、违反编写准则（如双重否定）、文化适应性等挑战，需要加强构念表达的准确性和测量特性的稳定性。

4.4 AIG 在心理健康测验中的应用

心理健康测验由于专业性强、伦理要求高，AIG 的应用一直较为有限。近期，王鹏等人（2024）将大语言模型与检索增强生成技术（RAG）相结合，通过构建心理咨询领域知识库并融合生成式 AI，实现了青少年心理危机评估量表的自动编制。

心理健康测验 AIG 面临专业知识库构建和伦理安全把控两大挑战，需要在技术创新的同时确保生成题目的专业性与伦理性。这要求加强心理学与人工智能的交叉融合，完善知识库建设和质量审查机制。

4.5 应用实践小结

通过对 AIG 技术在教育、认知、人格和心理健康测验中应用实践的系统考察，可以发现测验质量控制问题既有共性又有特异性。在共性方面，各类测验都面临生成内容的质量稳定性和专业知识表达的准确性等基础性挑战。但在具体表现上，每类测验有其独特的质量控制难点：教育测验强调专业知识表达的严谨性和题目难度的精确控制；认知测验注重认知成分的有效测量和心理测量特性的稳定性；人格测验关注构念效度和测量结构的稳定性；心理健康测验则需特别关注伦理安全和文化适应性。

这种共性与特性的并存表明，突破测量质量瓶颈需要采取分层策略：基础层面通过整合专业知识图谱和引入 RAG 机制来提升 LLMs 的知识支持，并建立评估机制以提升整体质量，领域层面则针对不同类型测验开发专门的技术解决方案。

5 题目自动生成的质量控制

在 AIG 技术追求生成效率的同时，如何保证生成题目的质量是不容忽视的核心问题。现有研究表明，质量控制不仅涉及多个关键维度的评估标准，还需要建立从技术控制到专家评审再到测量学评估的多层次保障体系。这种质量控制体系正在经历从单纯依赖人工评审向人机结合的智能化转型。本章将系统分析质量控制的关键维度和实现机制，重点探讨如何通过技术创

新提升质量控制的效率和可靠性。

5.1 质量控制的关键维度

题目自动生成的质量控制需要从内容准确性、语言规范性、文化公平性和测量学指标四个维度进行全面把控。

5.1.1 内容准确性

内容的准确性是题目质量的基础，这不仅体现在答案准确性和干扰项合理性上，还包括专业术语使用的规范性、概念间逻辑关系的严密性、确保题目能够准确反映待测构念，以及情境设置的真实性。Mitkov 等人（2023）强调在专业领域测验中，除了基本的语言质量标准，还需要建立特定的专业内容评估标准。研究表明，即使采用经过专业语料微调的 LLMs，在专业术语使用、临床情境描述等方面仍需要严格把控（von Davier, 2022）。同时，质量保障还需要特别关注知识更新和时效性问题，这要求建立定期更新和验证机制（von Davier, 2022）。

5.1.2 语言规范性

语言表达的规范性直接影响题目和测量效果。Smith 等人（2010）关注句子长度、句法结构清晰度等基本形式规范，以确保生成的题目适合语言学习者使用。而 Pino 等人（2008）则通过评估句子的复杂度、上下文清晰度、语法性和长度来识别合适的句子。从交际语言测试理论的角度，题目表达还应贴近实际使用场景，避免教科书式或脱离实际的表达方式。

5.1.3 文化公平性

文化公平性是确保测验适用性的重要维度。这要求在生成过程中注意避免文化偏见，确保题目内容对不同文化背景的测试者都具有同等的可理解性和适用性。例如，Duolingo 英语测试采用来自不同文化背景的专家团队进行评审，以识别并消除可能存在的文化偏见或敏感内容（LaFlair et al., 2023）。

5.1.4 心理测量学指标

心理测量学指标是评估题目质量的重要依据。在题目水平上，主要考察以下统计特征：(1) 难度指数应适中，避免过难或过易；（2）区分度指数需达到一定水平，确保题目能有效区分不同特质水平的被试；(3) 选项分析中各干扰项的选择率应相对均衡且具有合理的梯度分布，表明干扰项设计合理；（4）项目功能差异分析（DIF）应显示题目对不同群体（如性别、文化背景等）无测量偏差；（5）项目拟合度指标应表明题目符合测验的理论模型。在测验层面，则需关注信度指标（如重测信度、内部一致性系数）和效度证据（如内容效度、结构效度、效标关联效度），以及维度间的区分性。这些统计指标共同为题目质量提供了客观依据。

5.2 多层次质量控制体系

为确保自动生成题目的质量，需要建立从技术、专业到统计的多层次控制体系。

5.2.1 技术层：生成阶段的前置控制

在生成阶段就需要通过技术手段进行前置把控。侯凯（2020）探索了基于词语重复率的语料筛选和基于语义相似度的选项生成方法，以提升生成题目的规范性和自然度。高凯（2024）在数学题目生成中，通过引入知识图谱约束和双向训练策略，可将题目正确性提升至 68%。徐坚（2024）采用基于图注意力网络的模型，通过增强题目与原文的语义关联，提升了题目的语言流畅性和答案唯一性。通过应用 RAG 技术，构建专业知识库和设计合理的检索方案，可以在生成过程中对内容质量进行初步把控，显著提高生成内容的可信度和专业性（陈欣等, 2024; 王鹏等, 2024）。

5.2.2 专业层：专家评审机制

专家评审需要从多个维度进行系统评估。Götz 等人（2023）提出了双盲专家评审方法，组建包括比利时、德国、印度、荷兰等国家的专家团队，采用严格的一致性标准。每道题目至少需经过 3 位专家的内容审查和 2 位专家的公平性审查，只有获得全部专家一致认可的题目才能进入下一轮评估。这种方法显著提高了评审的客观性和题目质量（Hommel et al., 2022）。

5.2.3 统计层：心理测量学评估

统计评估是题目质量控制的最后一道关卡，主要包括两个阶段：预测试阶段，通过小规模施测收集初步数据，计算题目参数并筛选不合格题目；正式施测阶段，基于大样本数据进行系统评估。如果需要构建题库，还需要通过等值化技术确保不同批次生成题目的可比性。已有研究展示了心理测量学评估在实践中的应用。Rafatbakhsh 等人（2021）的分析表明自动生成的阅读理解题目具有良好的测量特性，题目难度系数在 0.60-0.65 之间，区分度达到 0.32-0.51；研究还采用探索性因素分析、验证性因素分析和探索性结构方程模型（ESEM）评估结构效度，结果显示生成题目具有良好的拟合效果。Sayin 等人（2024）通过 Rasch 模型分析，发现生成的题目具有良好的信度和局部独立性。

5.3 智能化质量控制的创新实践

随着人工智能技术的发展，质量控制正在向智能化方向演进。von Davier（2019）探索了基于机器学习的自动评分和质量预测方法，通过分析题目的语言特征和认知复杂度等多维指标，在生成阶段进行初步质量筛选，提高评估效率。Gorgun 等人（2024）基于 LLM 构建了题目质量的自动评估系统，通过 QLoRA 策略对 Llama 3 模型进行微调，并优化评估提示语模板，该方法在识别低质量题目方面的准确率达 82%，可作为高效的初筛工具。

尽管自动化评估技术不断进步，能够快速识别潜在的质量问题，但如 Attali 等人（2022）所强调，人工审核仍是确保高利害测验题目质量的必要环节。未来的质量控制将更多地采用人机协同的方式，通过智能化工具辅助专家决策，在提高效率的同时确保评估的可靠性。

6 问题与对策：AIG 技术的质量提升路径

AIG 技术正在从单一的题目生成工具向成熟的智能测验系统转变。这种转变体现在三个方面：技术方法从基于规则的模板发展为整合知识图谱和 RAG 的智能生成系统；应用场景从基础能力测验扩展到人格测验、心理健康测验等复杂领域；质量控制从依赖人工评审发展为结合智能化评估的多指标多层次保障体系。分析这一转变过程中的挑战及其解决方案，对推动 AIG 技术的科学发展具有重要启示意义。

面对这一发展趋势，AIG 技术需要在三个层面实现突破：在技术层面，探索融合认知理论与深度学习的生成框架，通过整合专业知识图谱和 RAG 机制提升 LLMs 的知识表达能力；在应用层面，构建支持实时更新的智能题库系统，丰富应用场景，实现基于学习分析的个性化测评；在实践层面，完善质量评估机制，建立从技术到测量的全流程质量保障。以下将具体分析 AIG 技术面临的五个核心挑战及其解决思路。

第一，主流技术路线均存在其固有局限。基于规则的方法需要平衡题目变化与认知成分的稳定性，过严的规则限制多样性，过松则影响可靠性。基于语料库的方法受限于语料获取难度和维护成本。基于 LLMs 的方法虽然具有强大的生成能力，但存在产生虚假内容、违反题目编写准则、难以精确控制题目难度等问题（von Davier, 2022），且主要局限于文本模态，难以处理图形推理、空间关系等多模态测验内容。针对这些技术瓶颈，未来需要探索多模态融合的技术路线，将认知理论、深度学习模型和多模态生成技术相结合，构建更全面的题目生成框架。

第二，专业领域知识表达对 AIG 技术提出了严格要求。即使采用最新的深度学习技术，在专业测验领域仍面临术语准确性、情境真实性和专业逻辑严密性等挑战（von Davier, 2018）。特别是在医学、法律等领域，不仅需要确保术语使用的恰当性，还要准确描述符合实践特点的专业场景，同时保证题目在理论框架内的合理性。以法律领域为例，约 43%的自动生成题目需要人工修改或无法使用（Mitkov et al., 2023）。解决这一问题需要通过构建专业知识图谱和采用 RAG 技术，提升模型的专业知识理解和表达能力。

第三是测验构念效度与公平性的系统保障问题。这一挑战在人格测验等心理测量领域尤为突出，需要考虑构念的跨文化稳定性和测量等价性。基于 LLMs 的自动生成系统可能因训练数据中的系统性偏差而产生文化或群体偏见。对此，需要建立包含专业性、公平性、文化适应性等多维度的评价标准，在 LLMs 的数据收集、模型训练和评估等环节建立科学的偏差检测和控制机制。

第四是 LLMs 大规模应用中的资源约束问题。在高利害测验场景中，训练专属大模型面临基础设施和技术团队的双重挑战。前者涉及高性能计算设备和存储系统的硬件投入，后者需要同时配备深度学习技术人员和测量学专家。对于资源受限的研究者，可以通过提示工程优化、知识图谱约束和 RAG 技术策略，提升现有模型的专业性和可靠性。

第五是语言测试中的语料自动筛选问题。阅读理解测试通常要求使用未经改写的真实语料，同时需要在话题适切性、语言真实性、认知负荷和文化敏感性等多维度上把控质量。实现自动筛选不仅需要评估表层语言特征，还需理解主题内容和语境适用性。针对这一挑战，可采取分层筛选策略：通过 NLP 技术评估语言特征，结合专家判断评估内容适切性，实现效率与质量的平衡。

参考文献

陈欣, 李蜜如, 周悦琦, 周同, 张峰. (2024). 基于大语言模型的试题自动生成路径研究. *中国考试,* (12), 39–48.

丁向民. (2008). *基于本体的多项选择题自动生成技术研究* (硕士学位论文). 南京航空航天大学.

高凯. (2024). *基于预训练模型的初等数学题目自动生成* (硕士学位论文). 电子科技大学, 成都.

侯凯. (2020). *小学语文自适应训练习题生成技术研究* (硕士学位论文). 湖南师范大学, 长沙.

李中权, 张厚粲. (2008). 计算机自动化项目生成概述. *心理科学进展, 16*(2), 348–352.

申弋斌. (2022). *数学应用问题的自动求解和文本生成* (博士学位论文). 华东师范大学, 上海.

王蕾. (2023). 人工智能生成内容技术在教育考试中应用探析. *中国考试,* (8), 19–27.

王鹏, 封迅, 康艳俊, 康春花. (2024). *青少年心理危机量表的项目自动生成: 基于生成式人工智能和 RAG 技术*. 中国科学院科技论文预发布平台. 2025-02-01 取自 https://chinaxiv.org/abs/202406.00361V1

肖文艳. (2019). *基于语料库的中小学英语词汇分析及试题自动生成研究* (博士学位论文). 江西师范大学, 南昌.

徐坚. (2024). 语义图支持的阅读理解型问题的自动生成. *智能系统学报, 19*(2), 420–428.

Agarwal, M., & Mannem, P. (2011). Automatic gap-fill question generation from text books. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 56–64). Association for Computational Linguistics.

Alsubait, T., Parsia, B., & Sattler, U. (2014). Generating multiple choice questions from ontologies: Lessons learnt. In C. M. Keet & V. Tamma (Eds.), *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)* (pp. 73–84). CEUR Workshop Proceedings.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255–291). Longman.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671–704.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Ball, F. (2002). Developing wordlists for BEC. *Research Notes*, *8*, 10–13.

Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes*, *26*, 2–4.

Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 661–674). Routledge.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, *2*(3), 1–29.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*(1), 9–48.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., . . . Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFFL 2000 spoken and written academic language corpus*. Educational Testing Service.

Bormuth, J. R. (1970). *On the theory of achievement test items*. University of Chicago Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (pp. 1877–1901). Curran Associates, Inc.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, *14*(2–4), 15–33.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In Barzilay, R., & Kan, M. Y. (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1342–1352).

Association for Computational Linguistics.
Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.
Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.
Embretson, S. E. (2005). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 251–267). Cambridge University Press.
Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, *50*(3), 328–344.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
Embretson, S. E., & Yang, X. (2006). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 747–768). Elsevier.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.
Fu, Y., Choe, E. M., Lim, H., & Choi, J. (2022). An evaluation of automatic item generation: A case study of weak theory approach. *Educational Measurement: Issues and Practice*, *41*(4), 10–22.
Gates, D. M. (2011). How to generate cloze questions from definitions: A syntactic approach. In *2011 AAAI Fall symposium series*. Association for the Advancement of Artificial Intelligence. Retrieved February 1, 2025, from https://cdn.aaai.org/ocs/4220/4220-17711-1-PB.pdf
Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.
Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*(3), 196–210.
Gorgun, G., & Bulut, O. (2024). Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice*. Advance online publication. https://doi.org/10.1111/emip.12663
Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning*, *2*(3), 210–224.
Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494–518.
Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, *18*(4), 277–296.
Guttman, L. (1959). Introduction to facet design and analysis. *Acta Psychologica*, *15*, 130–138.
Haladyna, T. M. (2013). Automatic item generation: A historical perspective. In M. J. Gierl & T. M.

Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13–25). Routledge.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2013). *Halliday's introduction to functional grammar* (4th ed.). Routledge.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema*, *16*(4), 696–701.

Hively, W. (1974). Introduction to domain-referenced testing. *Educational Technology*, *14*(6), 5–10.

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, *87*(2), 749–772.

Jiang, S., & Lee, J. S. (2017, September). Distractor generation for chinese fill-in-the-blank items. In Tetreault, J., Burstein, J., Leacock, C., & Yannakoudakis, H. (Eds.), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 143-148). Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2022). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Prentice-Hall.

Karamanis, N., Ha, L. A., & Mitkov, R. (2006). Generating multiple-choice test items from medical text: A pilot study. In N. Colineau, C. Paris, S. Wan, & R. Dale (Eds.), *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 111–113). Association for Computational Linguistics.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Proceedings of the 36th Conference on Neural Information Processing Systems* (pp. 22199−22213). Curran Associates, Inc.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204.

LaFlair, G., Yancey, K., Settles, B., & von Davier, A. A. (2023). Computational psychometrics for digital-first assessments: A blend of ML and psychometrics for item generation and scoring. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 107–123). Routledge.

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics*, *18*(2), 141–165.

Laverghetta, A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, . . . T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 414–428). Association for Computational Linguistics.

Liu, C. L., Wang, C. H., Gao, Z. M., & Huang, S. M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. In J. Burstein & C. Leacock (Eds.), *Proceedings of the Second Workshop on Building Educational Applications Using NLP* (pp. 1–8). Association for Computational Linguistics.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* (pp. 17–22). Association for Computational Linguistics.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*(2), 177–194.

Mitkov, R., Maslak, H., Ranasinghe, T., & Sosoni, V. (2023). Automatic generation of multiple-choice test items from paragraphs using deep neural networks. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 77–89). Routledge.

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. In M. B. Nunes & M. McPherson (Eds.), *Proceedings of the IADIS International Conference on e-Learning 2008* (pp. 427–434). International Association for Development of the Information Society.

Pino, J., Heilman, M., & Eskenazi, M. (2008). A selection strategy to improve cloze question quality. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 22−34). Institute of Intelligent Systems.

Rafatbakhsh, E., Ahmadi, A., Moloodi, A., & Mehrpour, S. (2021). Development and validation of an automatic item generation system for English idioms. *Educational Measurement: Issues and Practice*, *40*(2), 49–59.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1−67.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (pp. 33–58). Lawrence Erlbaum Associates.

Sakaguchi, K., Arase, Y., & Komachi, M. (2013). Discriminative approach to fill-in-the-blank quiz generation for language learners. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 238–242). Association for Computational Linguistics.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, *43*(1), 5–18.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498.

Shei, C. C. (2001). FollowYou!: An automatic language lesson generation system. *Computer Assisted Language Learning*, *14*(2), 129–144.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory

to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Lawrence Erlbaum Associates.

Smith, S., Avinesh, P. V. S., & Kilgarriff, A. (2010). Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of the ICON-2010: 8th International Conference on Natural Language Processing* (pp. 1–6). Macmillan Publishers.

Sumita, E., Sugaya, F., & Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In J. Burstein & C. Leacock (Eds.), *Proceedings of the Second Workshop on Building Educational Applications Using NLP* (pp. 61–68). Association for Computational Linguistics.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 5999–6009). Curran Associates Inc.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857.

von Davier, M. (2019). *Training Optimus Prime, MD: Generating medical certification items by fine-tuning OpenAI's GPT2 transformer model*. arXiv. Retrieved February 1, 2025, from https://doi.org/10.48550/arXiv.1908.08594

von Davier, M. (2023). Training Optimus Prime, MD: A case study of automated item generation using artificial intelligence–from fine-tuned GPT2 to GPT3 and beyond. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 90–106). Routledge.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems* (pp. 24824−24837). Curran Associates, Inc.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E. (2021). When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 159–168). Association for Computing Machinery.

Zhou, Q., & Huang, D. (2019). Towards generating math word problems from equations and topics. In *Proceedings of the 12th international conference on natural language generation* (pp. 494-503). Association for Computational Linguistics.

Automatic Item Generation: Technical Evolution and Quality Control

HAN Yuting[1,2,3], WANG Wenxuan[4], LIU Hongyun[5], YOU Xiaofeng[6,7]

([1] *Cognitive Science and Allied Health School, Beijing Language and Culture University, Beijing, 100083, China*)

([2] *Institute of Life and Health Sciences, Beijing Language and Culture University, Beijing, 100083, China*)

([3] *Key Laboratory of Language and Cognitive Science (Ministry of Education), Beijing Language and Culture University, Beijing, 100083, China*)

([4] *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077*)

([5] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([6] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

([7] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang 360111, China*)

Abstract: Automatic Item Generation (AIG) aims to address significant challenges in psychological and educational testing, including high development costs, low efficiency, maintenance difficulties, and security risks. However, ensuring item quality while improving efficiency remains a critical challenge. This review examines the theoretical foundations of AIG and traces its technical evolution from rule-based to data-driven approaches, systematically analyzes its applications across various test types, and investigates multi-level quality control mechanisms. Several improvement strategies are proposed: integrating cognitive theory with deep learning, enhancing domain knowledge through knowledge graphs and retrieval-augmented generation, optimizing prompt engineering, incorporating multimodal technologies, and implementing multi-level quality assessment. These innovations are intended to facilitate AIG's transformation from a simple generation tool into a sophisticated intelligent testing system, thereby enhancing the quality and reliability of automatically generated items.

Keywords： Automatic Item Generation; Quality Control; Psychometrics; Large Language Models